epiconcept

smart health

epitweetr version 0.1.24:
R package and interactive interface (Shiny app)

Francisco Orchard

October 2020

epitweetr objectives

The primary objective of epitweetr:

- to use Twitter **to detect early signals of potential threats/events** by topic and by geographical unit.

The secondary objective:

- to **enable the users through an interactive interface to explore** aggregated Twitter data by time, geographical location and topic

epiconcept smart health

Principles of using epitweetr

- Free as speech! Open source (EUPL-1.2) and available from CRAN
- Powerful :
    - Up to 1.5B tweets per year
    - Uses machine learning to detect geographical mentions on tweets
- Runs locally : It can run on a laptop. After downloading tweets all processing is local.
- Running continuously
    - Collect tweets, geolocate, detect alerts and send emails
    - It recovers automatically from downtimes
- Customisable :
    - Easily add your own topics
    - The R API allow to create your own reports

How does it work?

General architecture of epitweetr

- **3 main underlying processes**
  - **Tweet collection**
  - **Processing**
    - Obtaining location information (geolocalising)
    - Aggregating the data → counts of data by

      topic

      geographical unit

      top words within tweets

      specific users
  - **Signal detection and email alerts**

- **Frontend: Interactive application (Shiny app)**
  - Data visualisations
  - Configuration

  → Both can be done from the R package

General architecture of epitweetr

- **3 main underlying processes**
  - **Tweet collection**
  - **Processing**
    - Obtaining location information (geolocalising)
    - Aggregating the data → counts of data by
      topic
      geographical unit
      top words within tweets
      specific users
  - **Signal detection and email alerts**

Processes occur according to a set schedule, e.g. every 4 hours

- **Frontend: Interactive application (Shiny app)**
  - Data visualisations (dashboard)
  - Configuration

  → Both can be done from the R package

epiconcept smart health

General architecture of epitweetr

- **3 main underlying processes**
  - **Tweet collection**
  - **Processing**
    - Obtaining location information (geolocalising)
    - Aggregating the data → counts of data by

      topic

      geographical unit

      top words within tweets

      specific users
  - **Signal detection and email alerts**

Processes occur according to a set schedule, e.g. every 4 hours

Settings can be configured on the configuration page of the Shiny app

- **Frontend: Interactive application (Shiny app)**
  - Data visualisations (dashboard)
  - Configuration

  → Both can be done from the R package

epiconcept

smart health

epitweetr configuration tab

On the configuration tab, you can

- Check the status of process/pipelines
- Modify
  - topics and associated queries
  - languages for geolocation
  - the list of important users
  - country/region definitions
- Change general settings and settings for tweet collection, signal detection and alert generation

epitweetr configuration tab vs dashboard

- What are the differences between configuration tab and dashboard settings?

→ changes on the dashboard are exploratory

→ changes on the configuration tab modify the tool itself

- Changes on the configuration page can alter the tweet collection, detection process, alert process, etc.

Tweet collection

- epitweetr uses **Twitter Standard Search API**
  - Only limited tweets in past ~7 days available
  - Not exhaustive (focuses on relevance)

    Not all tweets are indexed

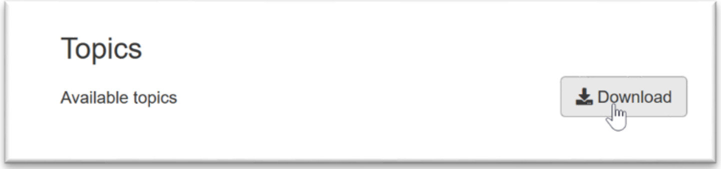
    Limited to 4.3M tweets per day
  - But it’s free!

  → Sufficient to meet objectives of tool

epiconcept
smart health

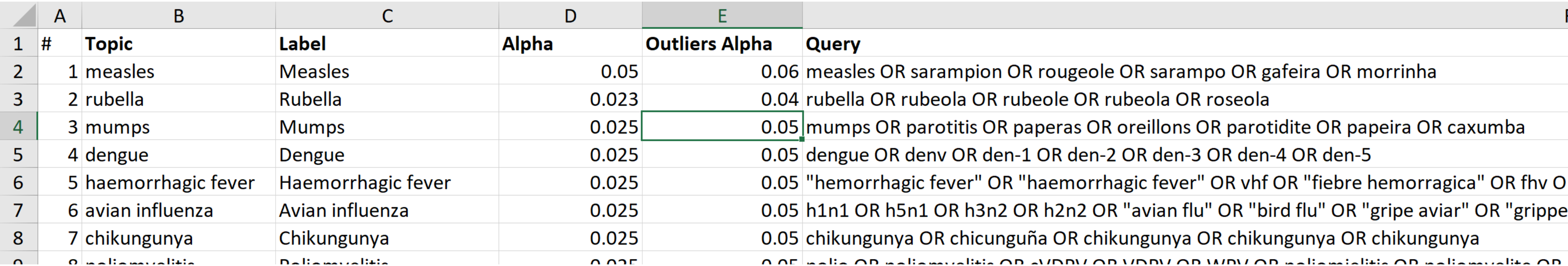
Tweet collection: topics

- Topics may be subject to changes (e.g. adding COVID-19 this year)
- Remember: Changes to queries do not affect historical data!
  - And never change a topic name, just the label
- Download the topics Excel spreadsheet to make modifications



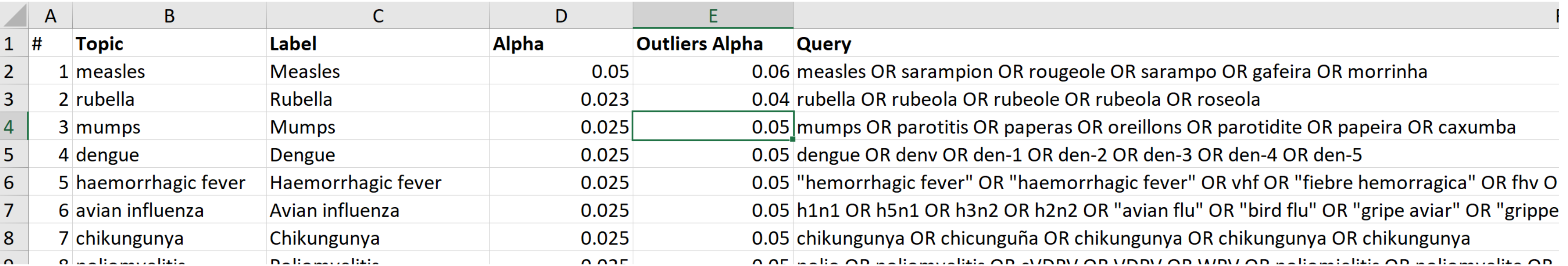
Topics

Available topics



	A	B	C	D	E	F
1	#	Topic	Label	Alpha	Outliers Alpha	Query
2	1	measles	Measles	0.05	0.06	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha
3	2	rubella	Rubella	0.023	0.04	rubella OR rubeola OR rubeole OR rubeola OR roseola
4	3	mumps	Mumps	0.025	0.05	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR caxumba
5	4	dengue	Dengue	0.025	0.05	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5
6	5	haemorrhagic fever	Haemorrhagic fever	0.025	0.05	"hemorrhagic fever" OR "haemorrhagic fever" OR vhf OR "fiebre hemorragica" OR fhv O
7	6	avian influenza	Avian influenza	0.025	0.05	h1n1 OR h5n1 OR h3n2 OR h2n2 OR "avian flu" OR "bird flu" OR "gripe aviar" OR "grippe
8	7	chikungunya	Chikungunya	0.025	0.05	chikungunya OR chicunguña OR chikungunya OR chikungunya OR chikungunya
9	8	poliomyelitis	Poliomyelitis	0.025	0.05	polio OR poliomyelitis OR cVDPV OR VDPV OR WPV OR poliomielitis OR poliomyelite OR

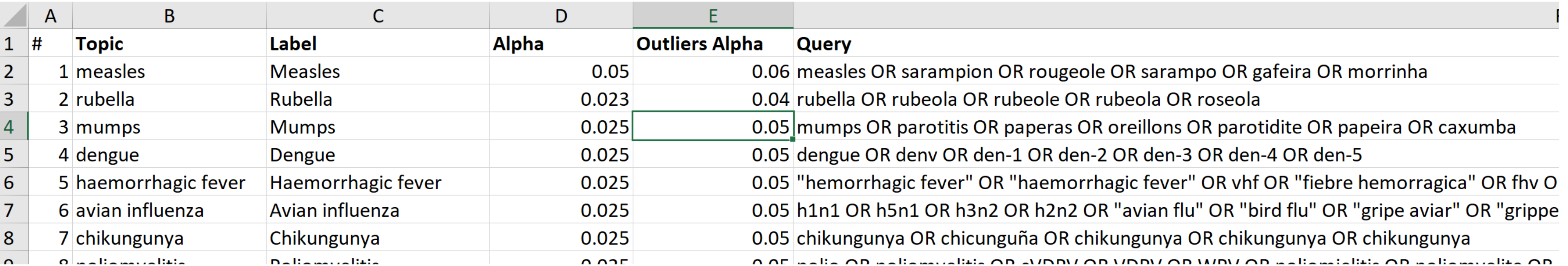
Tweet collection: topics



	A	B	C	D	E	F
1	#	**Topic**	**Label**	**Alpha**	**Outliers Alpha**	**Query**
2	1	measles	Measles	0.05	0.06	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha
3	2	rubella	Rubella	0.023	0.04	rubella OR rubeola OR rubeole OR rubeola OR roseola
4	3	mumps	Mumps	0.025	0.05	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR caxumba
5	4	dengue	Dengue	0.025	0.05	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5
6	5	haemorrhagic fever	Haemorrhagic fever	0.025	0.05	"hemorrhagic fever" OR "haemorrhagic fever" OR vhf OR "fiebre hemorragica" OR fhv O
7	6	avian influenza	Avian influenza	0.025	0.05	h1n1 OR h5n1 OR h3n2 OR h2n2 OR "avian flu" OR "bird flu" OR "gripe aviar" OR "grippe
8	7	chikungunya	Chikungunya	0.025	0.05	chikungunya OR chicunguña OR chikungunya OR chikungunya OR chikungunya
9	8	poliomyelitis	Poliomyelitis	0.025	0.05	polio OR poliomyelitis OR cVDPV OR VDPV OR WPV OR poliomielitis OR poliomyelite OR

- You can add a topic → add a new line with Topic name and Label (the label is what appears in the dashboard dropdown menu)
- Alpha: Signal detection false positive rate: with a higher value, potentially fewer “true” signals will be missed
- Outliers alpha: Outliers false positive rate: Threshold to determine which outliers to downweight: with a higher value, potentially more data points downweighted

Tweet collection: topics

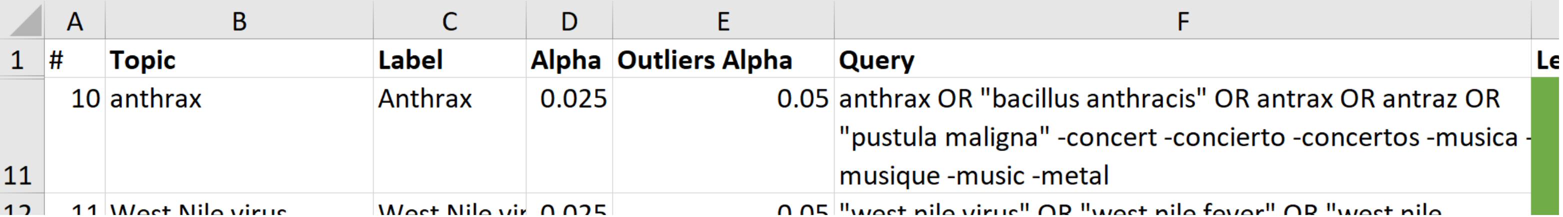


	A	B	C	D	E	F
1	#	**Topic**	**Label**	**Alpha**	**Outliers Alpha**	**Query**
2	1	measles	Measles	0.05	0.06	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha
3	2	rubella	Rubella	0.023	0.04	rubella OR rubeola OR rubeole OR rubeola OR roseola
4	3	mumps	Mumps	0.025	0.05	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR caxumba
5	4	dengue	Dengue	0.025	0.05	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5
6	5	haemorrhagic fever	Haemorrhagic fever	0.025	0.05	"hemorrhagic fever" OR "haemorrhagic fever" OR vhf OR "fiebre hemorragica" OR fhv O
7	6	avian influenza	Avian influenza	0.025	0.05	h1n1 OR h5n1 OR h3n2 OR h2n2 OR "avian flu" OR "bird flu" OR "gripe aviar" OR "grippe
8	7	chikungunya	Chikungunya	0.025	0.05	chikungunya OR chicunguña OR chikungunya OR chikungunya OR chikungunya
9	8	poliomyelitis	Poliomyelitis	0.025	0.05	polio OR poliomyelitis OR cVDPV OR VDPV OR WPV OR poliomielitis OR poliomyelite OR

- Query best practice: Limit your searches to 10 keywords and operators
- A space indicates an "AND": hemorrhagic fever
  - Returns tweets with "hemorrhagic" and "fever" (but not necessarily next to each other)
- Quotation marks indicate an exact phrase: "hemorrhagic fever"
  - Returns tweets with "hemorrhagic fever" in them

Tweet collection: topics

- OR: you can look for more than one keyword: “hemorrhagic” OR “fever”
  - Returns tweets with at least one of “hemorragic” or “fever”

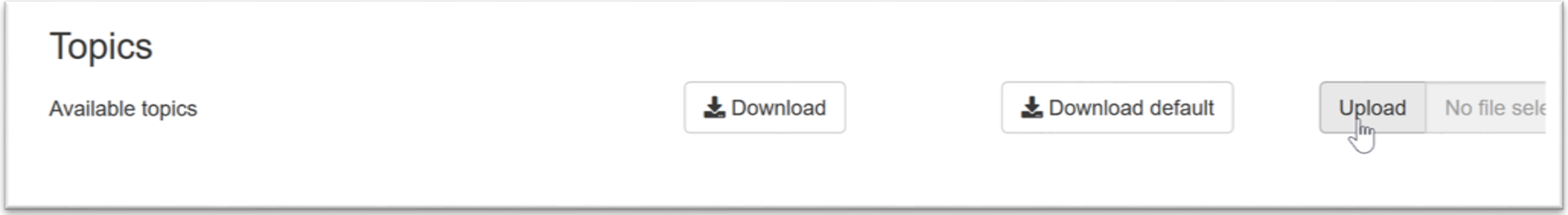


	A	B	C	D	E	F	
1	#	**Topic**	**Label**	**Alpha**	**Outliers Alpha**	**Query**	**Le**
11	10	anthrax	Anthrax	0.025	0.05	anthrax OR "bacillus anthracis" OR antrax OR antraz OR "pustula maligna" -concert -concierto -concertos -musica -musique -music -metal	🟩
12	11	West Nile virus	West Nile vir	0.025	0.05	"west nile virus" OR "west nile fever" OR "west nile	🟩

- A dash before the keyword (no space) means NOT
  - anthrax –metal
  - Returns tweets containing anthrax but not metal
- A query can have maximum 500 characters

Tweet collection: topics

- Save your changes and upload:



Topics

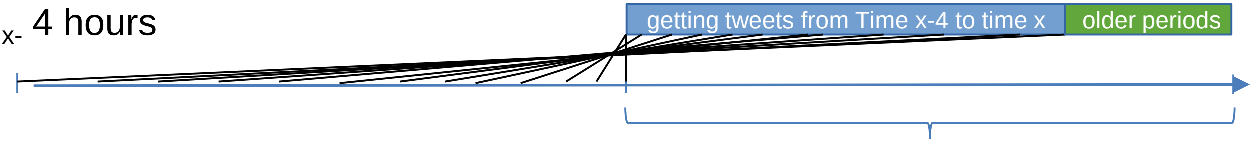
Available topics

- Made a mistake?
- Download the default!

Tweet collection: process

- Queries sent to Twitter API within a regular schedule (e.g. 4 hours)
  → Collection of tweets from:
  current time to (current time – schedule)

  Assuming a 4 hour schedule:



Time$_x$

Time$_{x-}$ 4 hours

Time$_x$ + 4 hours

Arrow represents real time elapsing

epitweetr collects tweets for a given request from current time going backwards to (Time$_x$ – 4 hours)

- If all tweets are collected early within a schedule, epitweetr will use remaining time to help collect outstanding tweets from previous time periods

epiconcept
smart health

The tweets are collected and stored

What do we do next?

→ we need to know which country the topic is about

e.g.: Ebola in DRC

Measles in the UK

Campylobacter in France

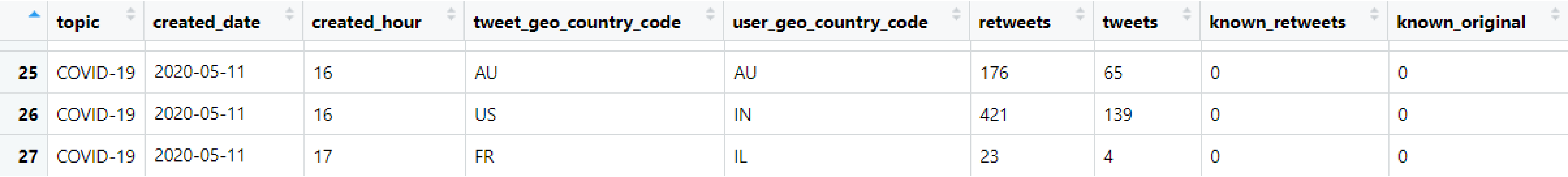
Processing: geolocation

- Parallel to tweet collection, tweets are geolocated using the same schedule
- Geographical information in a tweet is available
    - In the **tweet text** (of current or retweeted tweet)
        - This is considered the more “valid” geolocation
    - In **user information** (biography, user location at time of tweeting)
- **epitweetr stores both tweet and user geographical information** separately,
  as a locality, country and region name, e.g. Conakry, Guinea, Africa and a longitude and latitude
- No information is stored if no geographical information available in the tweet text and user information, respectively
- epitweetr uses machine learning to find geographical info within tweet text
    - Using the (free) geonames database

Processing: Aggregation of data

So far: tweets collected and geographical information stored, next:

- Data are aggregated into 3 R datasets (Rds) to be used for visualisation (Shiny app) and alert detection
- country_counts.Rds:

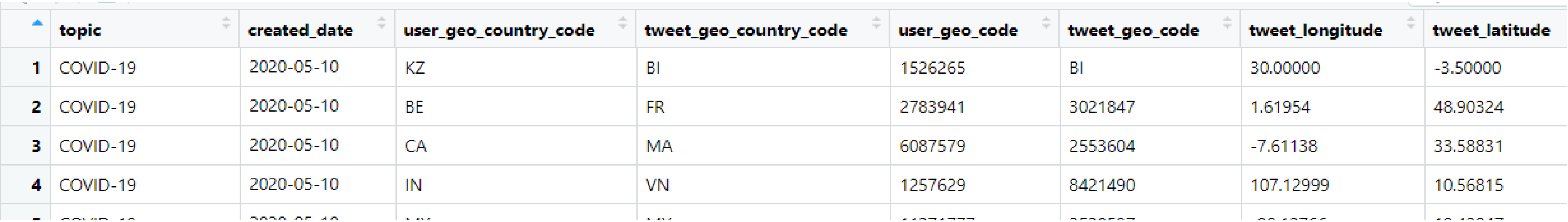


	topic	created_date	created_hour	tweet_geo_country_code	user_geo_country_code	retweets	tweets	known_retweets	known_original
25	COVID-19	2020-05-11	16	AU	AU	176	65	0	0
26	COVID-19	2020-05-11	16	US	IN	421	139	0	0
27	COVID-19	2020-05-11	17	FR	IL	23	4	0	0

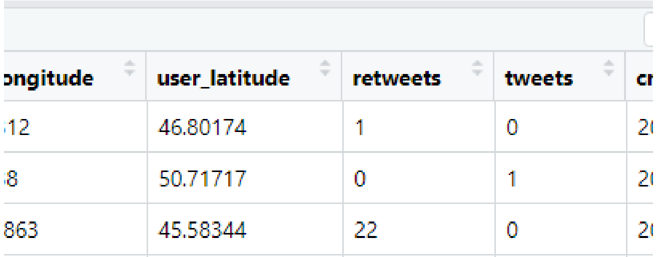
Number of tweets from specific users
The users can be modified via the configuration page

Processing: Aggregation of data

- geolocated.Rds, used for the map:



	topic	created_date	user_geo_country_code	tweet_geo_country_code	user_geo_code	tweet_geo_code	tweet_longitude	tweet_latitude
1	COVID-19	2020-05-10	KZ	BI	1526265	BI	30.00000	-3.50000
2	COVID-19	2020-05-10	BE	FR	2783941	3021847	1.61954	48.90324
3	COVID-19	2020-05-10	CA	MA	6087579	2553604	-7.61138	33.58831
4	COVID-19	2020-05-10	IN	VN	1257629	8421490	107.12999	10.56815
5	COVID-19	2020-05-10	MX	MX	[illegible]	[illegible]	[illegible]	[illegible]

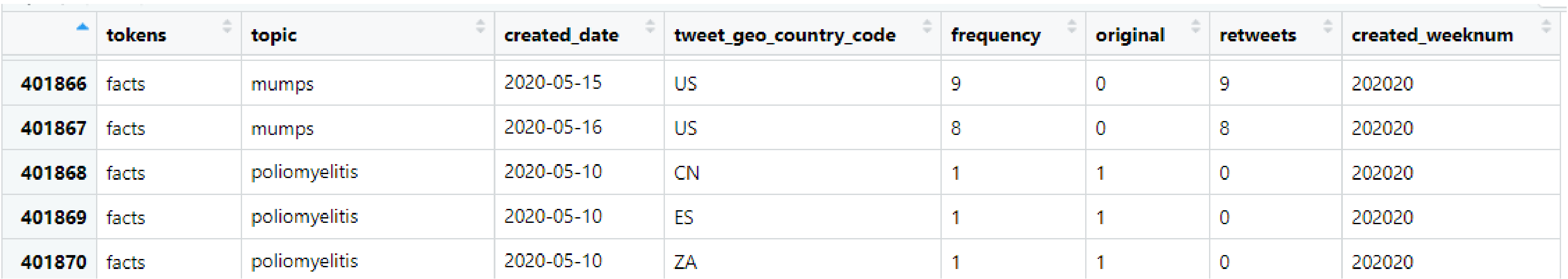


ongitude	user_latitude	retweets	tweets	cr
12	46.80174	1	0	2
8	50.71717	0	1	2
863	45.58344	22	0	2

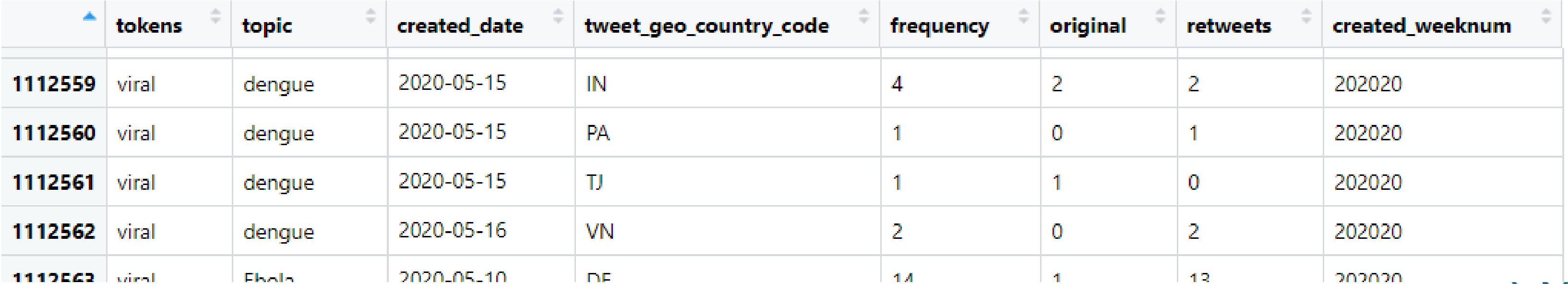
epiconcept
smart health

Processing: Aggregation of data

- topwords.Rds



	tokens	topic	created_date	tweet_geo_country_code	frequency	original	retweets	created_weeknum
401866	facts	mumps	2020-05-15	US	9	0	9	202020
401867	facts	mumps	2020-05-16	US	8	0	8	202020
401868	facts	poliomyelitis	2020-05-10	CN	1	1	0	202020
401869	facts	poliomyelitis	2020-05-10	ES	1	1	0	202020
401870	facts	poliomyelitis	2020-05-10	ZA	1	1	0	202020

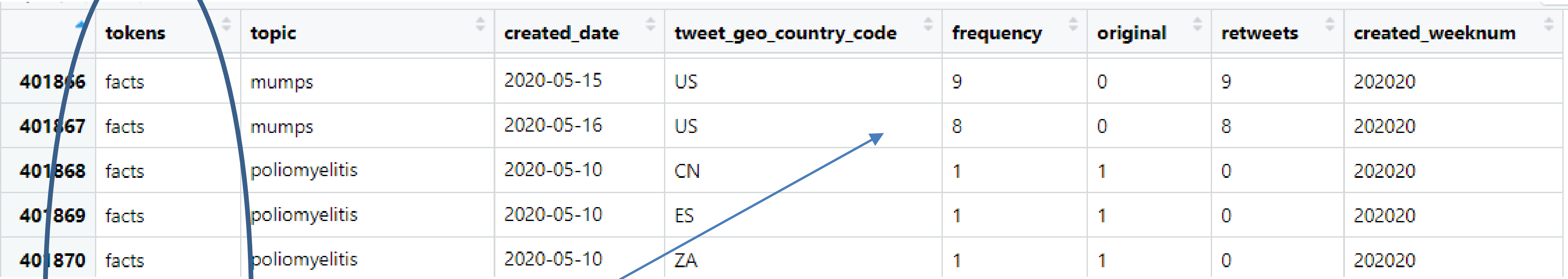


	tokens	topic	created_date	tweet_geo_country_code	frequency	original	retweets	created_weeknum
1112559	viral	dengue	2020-05-15	IN	4	2	2	202020
1112560	viral	dengue	2020-05-15	PA	1	0	1	202020
1112561	viral	dengue	2020-05-15	TJ	1	1	0	202020
1112562	viral	dengue	2020-05-16	VN	2	0	2	202020
1112563	viral	Ebola	2020-05-10	DE	14	1	13	202020

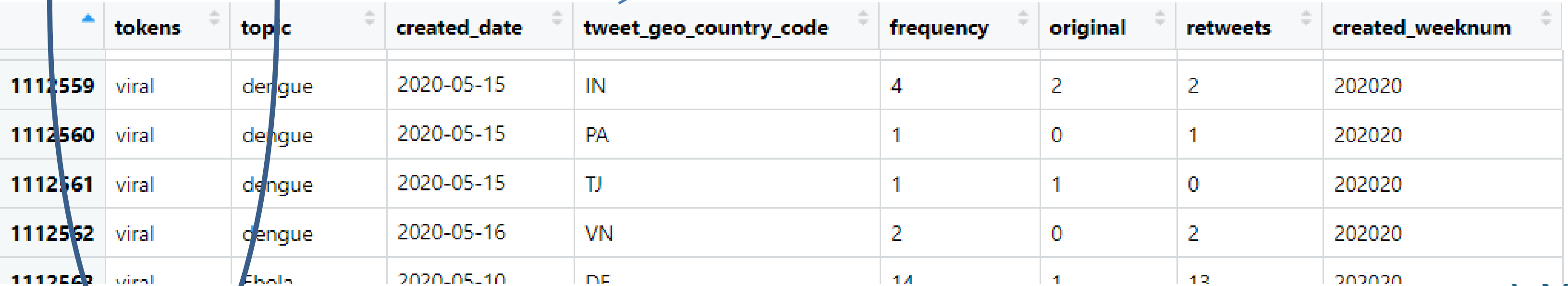
Processing: Aggregation of data

- topwords.Rds

Key words are collected for each topic



	tokens	topic	created_date	tweet_geo_country_code	frequency	original	retweets	created_weeknum
401866	facts	mumps	2020-05-15	US	9	0	9	202020
401867	facts	mumps	2020-05-16	US	8	0	8	202020
401868	facts	poliomyelitis	2020-05-10	CN	1	1	0	202020
401869	facts	poliomyelitis	2020-05-10	ES	1	1	0	202020
401870	facts	poliomyelitis	2020-05-10	ZA	1	1	0	202020



	tokens	topic	created_date	tweet_geo_country_code	frequency	original	retweets	created_weeknum
1112559	viral	dengue	2020-05-15	IN	4	2	2	202020
1112560	viral	dengue	2020-05-15	PA	1	0	1	202020
1112561	viral	dengue	2020-05-15	TJ	1	1	0	202020
1112562	viral	dengue	2020-05-16	VN	2	0	2	202020
1112563	viral	Ebola	2020-05-10	DE	14	1	13	202020

I have collected tweets and assigned a location to (some of) them. I’ve stored them in a nice format. How can I tell if something out of the ordinary is happening?

epiconcept

smart health

Signal detection

- epitweetr determines if number of tweets by topic/location exceeds the expected
- Uses a modified EARS algorithm[1] (Early Aberration Reporting System), part of surveillance R package[2]

$$\bar{y}_t + t_{1-\alpha}(k-1) \cdot s_t \cdot \sqrt{1+\frac{1}{k}},$$

y_t = mean
k = number of days in baseline
$s_{t =}$ standard deviation
$t_{1-\alpha}(k-1)$ denotes the$1-\alpha$ quantile of the t-distribution with $k-1$ degrees of freedom;

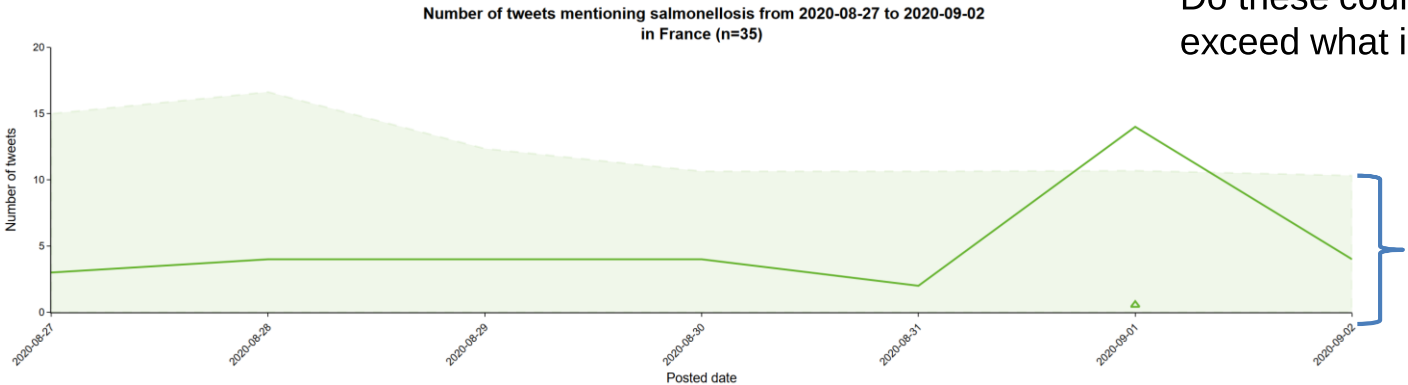
- Counts for a 24 hour window are checked to see if they exceed a threshold, based on data from the past 7 days
- epitweetr downweights previous outliers, in order not to miss a signal
- Signal generated if the threshold is exceeded

[1] Fricker et al, 2008, "Comparing Syndromic Surveillance Detection Methods: EARS' versus a CUSUM-Based Methodology." *Statistics in Medicine*

[2] Salmon et al, 2016, "Monitoring count time series in R: Aberration detection in public health surveillance" *Journal of Statistical Software*

Signal detection

- epitweetr determines if number of tweets by topic/location exceeds the expected
- Uses a modified EARS algorithm[1] (Early Aberration Reporting System), part of surveillance R package[2]



Do these counts exceed what is expected?

Threshold

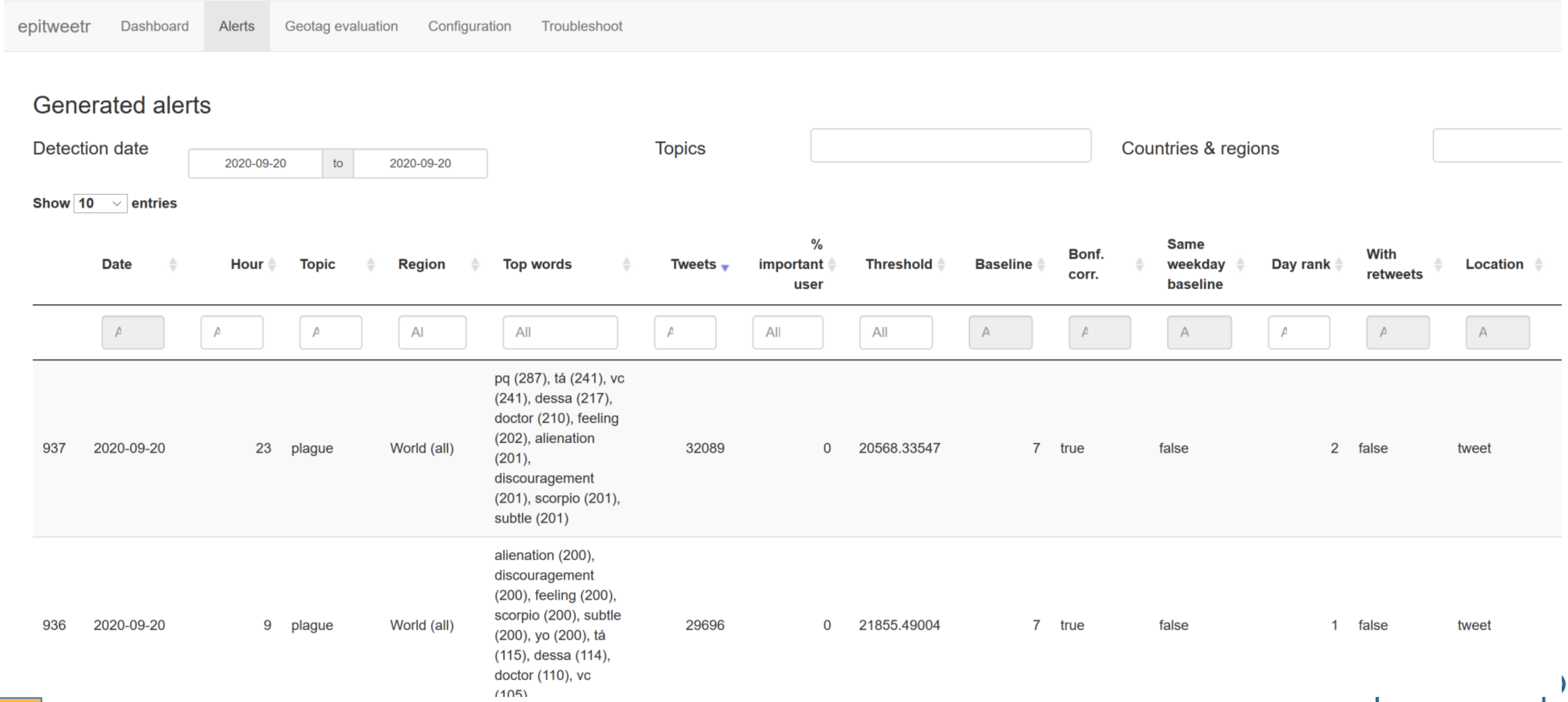
- epitweetr downweights previous outliers, in order not to miss a signal
- Signal generated if the threshold is exceeded

[1] Fricker et al, 2008, “Comparing Syndromic Surveillance Detection Methods: EARS’ versus a CUSUM-Based Methodology.” *Statistics in Medicine*
[2] Salmon et al, 2016, “Monitoring count time series in R: Aberration detection in public health surveillance” *Journal of Statistical Software*

Signal detection

- The signal detection is an ongoing process
- Every x hours, depending on your settings (e.g. every 4 hours), email alerts are sent summarising the signals, including:
  - Date and time slot of the signals
  - Locations where signals were detected
  - Number of tweets and percentage of excess tweets (by time and location)
  - Number of tweets from trusted users (by time and location)
  - Most frequent words (by time and location)
- Alerts sent via email; also available on Alerts tab on Shiny app

Signal detection: Alerts tab



Generated alerts

Detection date 2020-09-20 to 2020-09-20

Topics

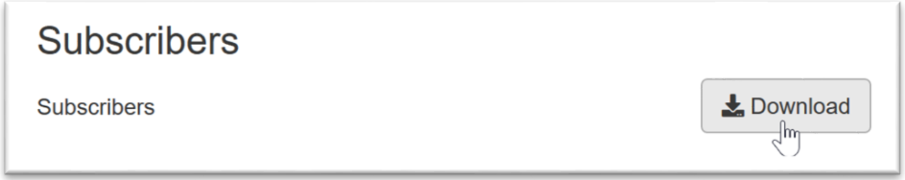
Countries & regions

Show 10 entries

	Date	Hour	Topic	Region	Top words	Tweets	% important user	Threshold	Baseline	Bonf. corr.	Same weekday baseline	Day rank	With retweets	Location
	A	A	A	Al	All	A	All	All	A	A	A	A	A	A
937	2020-09-20	23	plague	World (all)	pq (287), tá (241), vc (241), dessa (217), doctor (210), feeling (202), alienation (201), discouragement (201), scorpio (201), subtle (201)	32089	0	20568.33547	7	true	false	2	false	tweet
936	2020-09-20	9	plague	World (all)	alienation (200), discouragement (200), feeling (200), scorpio (200), subtle (200), yo (200), tá (115), dessa (114), doctor (110), vc (105)	29696	0	21855.49004	7	true	false	1	false	tweet

Signal detection: Subscribers of email alerts

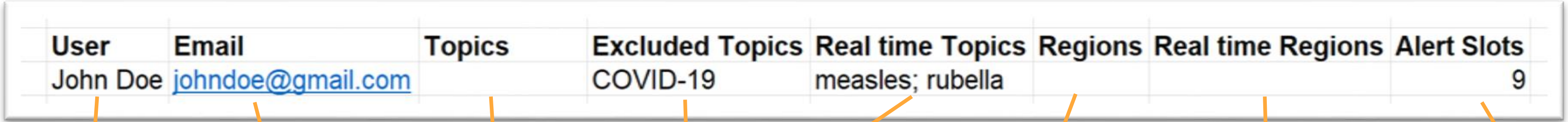
- Subscribers section on the configuration tab gives information on email alert and recipient properties
- Subscribers can receive
  - Real-time alerts (as soon as alert is available in epitweetr)
  - Scheduled alerts (e.g. 1 or 2 times a day)
- Download the Excel spreadsheet to modify Subscribers settings:



Subscribers

Subscribers

Signal detection: Subscribers of email alerts



User	Email	Topics	Excluded Topics	Real time Topics	Regions	Real time Regions	Alert Slots
John Doe	johndoe@gmail.com		COVID-19	measles; rubella			9

Name of email recipient

Recipient’s email address

Topics included in scheduled emails (blank = all)

Topics excluded from all emails

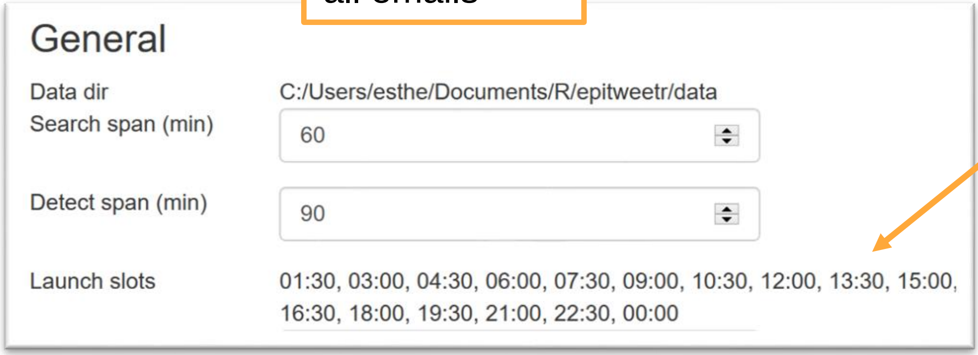
Topics received in “real-time” emails

Regions included in scheduled alerts

Regions included in real-time alerts

Detection loop slots after which subscriber receives emails; if empty, all alerts will be in real-time

Separate topics, regions and alert slots with a semi-colon “;”



General

Data dir C:/Users/esthe/Documents/R/epitweetr/data

Search span (min) 60

Detect span (min) 90

Launch slots 01:30, 03:00, 04:30, 06:00, 07:30, 09:00, 10:30, 12:00, 13:30, 15:00, 16:30, 18:00, 19:30, 21:00, 22:30, 00:00

How can I decide whether it is signal of a public health threat of interest?

epiconcept

smart health